**Title:-** Uncovering Patterns in Hotel Booking Data for Operational Efficiency and Revenue Growth

**Introduction :-**

This report explores a Resort Hotel's booking data. The main goal is to find important patterns and trends to help the hotel make smarter decisions. This case study conducts an Exploratory Data Analysis (EDA) on hotel booking records for both a Resort and a City Hotel. The aim is to discover operational patterns, guest behavior trends, and factors influencing revenue, ultimately to improve hotel management decisions

**Objectives**

The key objectives of this Exploratory Data Analysis project are:

* + Clean and preprocess the dataset for reliable analysis.
  + Explore trends in lead time, customer demographics, and booking behavior.
  + Investigate relationships between features and key metrics like ADR.
  + Conduct hypothesis tests to validate business assumptions.

**Project Summary:-**

This project involves the exploratory analysis of a dataset containing hotel booking information. The raw dataset initially comprised 119,390 entries across 32 features. The initial steps focused on data cleaning, where duplicate records were removed to ensure data integrity, resulting in a refined dataset of 87,396 unique bookings. Missing values were identified across several columns, notably company, agent, and country

**Data Cleaning and Preprocessing:-**

Data cleaning and preprocessing are crucial steps to ensure the reliability and accuracy of any subsequent analysis. The following steps were performed:

* **Library Imports and Data Loading:** Essential libraries like pandas, numpy, matplotlib, and seaborn were imported. The hotel\_bookings.csv dataset was then loaded.
* **Dataset Inspection:** The raw dataset was found to have 119,390 rows and 32 columns.
* **Duplicate Handling:** A significant number of duplicate records, precisely 31,994, were identified and subsequently removed. This process resulted in a refined dataset containing 87,396 unique booking records.
* **Missing Value Management:**
* Missing values in the children, country, and agent columns were addressed by filling them with the mode (most frequent value) of each respective column.
* The company column was entirely removed from the dataset due to an overwhelming proportion of missing values (over 93%).

**Data Type and Structure Verification:** Data types were checked and adjusted as needed, ensuring that both categorical and numerical variables were correctly handled for subsequent analysis. A bar plot was used to visualize the extent of missing values in each column, aiding in decision-making regarding their treatment.

**Outlier Treatment:-**

Outlier were found in the **lead time** (how far in advance someone booked) and **ADR** (average daily rate).

* **Using Boxplots :-** These are like special charts that show us where most of our data lies and clearly highlight any points that fall far outside the normal range.
* **Skewness Distribution Plots:-** These plots helped us see the overall shape of our data and if it was heavily weighted to one side, which often indicates the presence of outliers.

# Applied IQR-Based Capping: Used the Interquartile Range (IQR) method to cap extreme values in lead\_time and adr, By doing this, we minimized the impact of these extreme outliers, making our data much more reliable and accurate for any further analysis.

* Identified Variable Types:-

The project involved meticulous steps for feature engineering and data wrangling:

* Identifying categorical variables (e.g., hotel, customer\_type, reservation\_status)
* discrete numerical variables (e.g., is\_canceled, agent, adr)
* continuous numerical variables (e.g., lead\_time, adr)
* Analyzed Discrete Variables:-

Discrete numerical variables, like is\_canceled, agent, and adr, were examined separately due to their restricted number of unique values, despite their numerical nature.

* Handling Categorical Variables:-

Categorical features were processed effectively to allow for grouping, visualization, and the calculation of summary statistics.

* Continuous Variables:-

Continuous variables such as lead\_time and adr were singled out for specific outlier handling and subsequent statistical evaluations.

# Diagram Used:

# Value Counts per Variable: Displayed the number of unique values in each column to support the classification into categorical, discrete, and continuous types.

**Exploratory Data Analysis (EDA)**

The EDA phase involved key visual analyses to understand the dataset.

* **Univarite Analysis:-**

Histogram were created for :-

adr(Average Daily Rate)

lead\_time

customer\_type

market\_segment

along with other categorical and numerical variables, to observe their individual distributions.

* **Bivariate Analysis:**

Boxplots were used to compare the ADR across different market segments.

Heatmap of correlation matrix to identify relationships among numerical variables and detect multicollinearity

* **Time Series Analysis:** Booking trends were examined on a monthly and yearly basis to uncover seasonality and demand patterns

**Insights Gained:**

* Online Travel Agencies (OTA) dominate the booking volume across both hotel types.
* ADR (Average Daily Rate) shows noticeable variation across distribution channels.

Guests with higher lead times tend to make more booking changes, indicating longer decision windows.

**Correlation Analysis**

* adr(Average Daily Rate) shows moderate positive correlation with:-

total\_of\_special\_requests

lead\_time

* Weak correlations observed between:

booking\_changes and adr

previous\_cancellations and is\_canceled

These results suggest that while some variables are mildly related to pricing or behavior, others may have minimal impact.

Diagram Used:

* Heatmap of the Correlation Matrix

Highlighted stronger and weaker relationships among numerical features.

Useful for identifying patterns and potential multicollinearity issues in the data.

**Hypothesis Testing**

Tests Performed:

1. **ADR: OTA vs Direct Bookings**

Null Hypothesis (H₀): There is no significant difference in ADR between Online Travel Agencies and Direct bookings.

Result: Rejected H₀ — A significant difference was found, indicating that booking channel influences room pricing.

1. **Room Upgrades vs Lead Time**

Null Hypothesis (H₀): Room upgrades are independent of the lead time.

Result: Failed to reject H₀ — No strong statistical evidence to suggest a relationship.

1. **Stay Duration vs Customer Type**

Null Hypothesis (H₀): There is no significant difference in stay duration across different customer types.

Result: Rejected H₀ — Significant variation in stay durations among customer categories (e.g., transient vs. contract).

**Diagram Used:**

* Boxplots – To compare distributions of ADR and stay durations across groups.
* Grouped Bar Charts – To visualize categorical comparisons and summarize group-based differences effectively.

**Key Findings**

Hotel and Booking Trends

Here are the key trends observed in hotel operations and bookings:

* Hotel Type Preference: City hotels are significantly more popular, receiving 61.07% of all bookings compared to resort hotels.
* Average Daily Rate (ADR) Drivers: The ADR is highest for guests who:

Make special requests.

Book with longer lead times.

* Cancellation Rate: Approximately one in every four reservations (25%) results in a cancellation.
* Most Favored Meal Plan: BB (Bed & Breakfast) is the preferred meal plan among guests.
* Primary Booking Platform: Bookings are predominantly made through the Online Travel platform.
* Leading Distribution Channel: TA/TO (Travel Agents/Tour Operators) are the leading distribution channel for bookings.
* Busiest Months: September and October are the busiest months for bookings.
* Minimal Wait Times: July experiences the shortest wait times for reservations.
* Room Assignment Mismatches: While room assignment mismatches are common, they do not significantly impact the Average Daily Rate (ADR) or lead to cancellations.

**Challenges Faced**

* + High dimensional dataset with categorical and numerical mix.
  + Null values and duplicate records.
  + Outlier skew affecting distribution shapes.
  + Complex feature relationships.
  + The data contained a large number of duplicates.
  + The improper data type format was used for the data.
  + It was challenging to select the best visualization techniques.
  + The dataset contained a large number of null values.

**Solution to Business Objective**

**Here's a well-formatted breakdown of the proposed strategies for hotel optimization:**

Strategic Recommendations for Hotel Optimization

**1. Optimize Resort Hotel Bookings**

Since city hotels receive more bookings, we should introduce targeted packages and seasonal promotions specifically designed to attract more customers to our resort locations.

**2. Diversify Meal Plan Preferences**

While BB (Bed & Breakfast) is the most requested meal type, we must maintain its high quality. To diversify preferences, ease kitchen load, and encourage variety, we should offer discounts on other meal plans.

**3. Streamline Online Booking Channels**

Given that most bookings originate from online platforms, we can improve efficiency by reducing or eliminating underperforming segments like complementary and aviation, which contribute minimally to overall bookings.

**4. Enhance Distribution Channel Strategy**

* Invest more in our top-performing channels: TA/TO (Travel Agents/Tour Operators) and Corporate channels, as they generate the highest number of bookings.
* Consider phasing out the GDS channel due to its very low activity.

**5. Boost Repeat Bookings**

With only 3.86% of guests being repeat customers, we need to implement strategies to improve retention:

* Offer loyalty programs or repeat booking incentives.
* Personalize marketing efforts based on guest history.
* Identify and target priority customers for specific retention initiatives.

**6. Rationalize Room Type Management**

* Maintain and enhance Room Types A and D, as they are the most preferred by guests.
* Promote Room Types E, F, and G through discounts to balance demand across our inventory.
* To reduce operational costs, eliminate Room Types B, C, H, and L, which are less frequently booked.

**7. Encourage Advance Deposits**

Guests currently avoid pre-deposit bookings. We need to educate guests on the benefits of advance deposits and actively promote them to:

* Accelerate revenue recognition.
* Reduce cancellations and no-shows.

**8. Optimize Parking Space Allocation**

Since spaces for 3 and 8 cars are rarely booked, we should limit our offerings to 1 or 2 spaces to optimize resources and maximize space efficiency.

**9. Improve Room Assignment Accuracy**

With 15% of guests not receiving their reserved room, it's crucial to ensure better room allocation accuracy to significantly improve customer satisfaction.

**10. Mitigate High Cancellation Rate**

Given the 25% cancellation rate, we should implement several measures:

* Establish a flexible yet firm cancellation policy.
* Offer discounts for non-refundable bookings.
* Send reminder notifications to guests to help minimize cancellations.

**11. Promote Group Bookings**

As most bookings are currently for two people, we can increase revenue by actively promoting family and group stays. This can be achieved by offering:

* Bundle deals.
* Group discounts.
* Event-specific promotions.

# **Conclusion**

This Exploratory Data Analysis (EDA) offered insights into how guests book and how hotels operate. By cleaning the data, visualizing important details, and testing ideas using hypotheses, the study identified key areas for improvement: adjusting pricing based on customer groups and booking methods, better managing room assignments, and focusing on guests who bring in more revenue (higher ADR).

**Tools & Libraries Used**

**Programming Language:**

Python

**Python Libraries:**

Pandas – for data manipulation and analysis

NumPy – for numerical operations

Matplotlib & Seaborn – for data visualization

Scipy – for statistical testing and hypothesis analysis

**Development Environment:**

**Jupyter Notebook** – for interactive coding, analysis, and visual documentation

# **References**

UCI Hotel Booking Demand Dataset

Assignment brief from CDAC Mumbai (AAS Case Study Feb 2025)

**Submitted by :** Mrunali Gaikwad